Customer Shopping Behaviour Analysis

1. Project Overview

This project analyses customer shopping behaviour using transactional data from 3900 purchases across various product categories. The goal is to uncover insights into spreading patterns, customer segments, product preference and subscription behaviour to guide strategic business decisions.

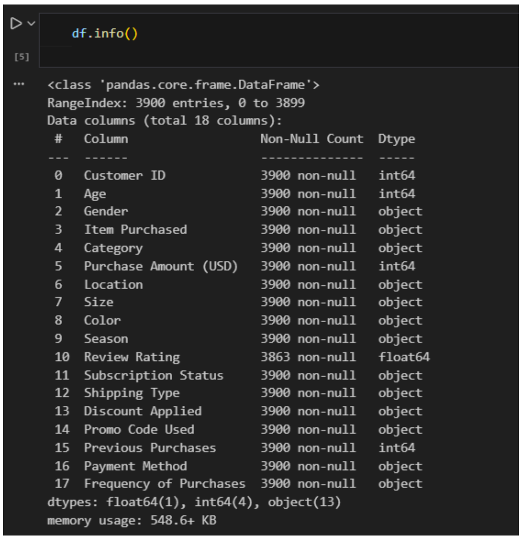
2. Dataset Summary

- **Rows:** 3900
- **Columns:**18
- **Key Features:**
  - Customer demographics (Age, Gender, Locations, Subscription status)
  - Purchase Details (Item Purchased, Categories, Purchase Amount, Season, Size, Colour)
  - Shopping Behaviour (Discount Applied, Promo Code Used, Previous Purchase, Frequency of Purchase, Review Rating, Shipping Type)
- **Missing Data:** 37 values in Review Rating column

3. Exploratory Data Analysis Using Python

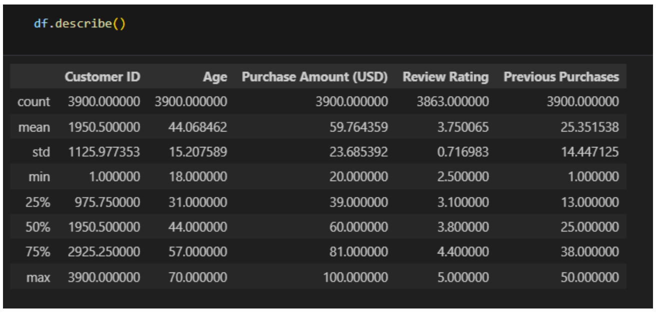
We can begin with data preparation and cleaning in python:

- **Data Loading:** Imported the data using pandas.
- **Initial Exploration:** Used df.info() to check structure and df.descrbe() for summary statistics.



```
df.info()
[5]

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Customer ID             3900 non-null   int64  
 1   Age                     3900 non-null   int64  
 2   Gender                  3900 non-null   object 
 3   Item Purchased          3900 non-null   object 
 4   Category                3900 non-null   object 
 5   Purchase Amount (USD)   3900 non-null   int64  
 6   Location                3900 non-null   object 
 7   Size                    3900 non-null   object 
 8   Color                   3900 non-null   object 
 9   Season                  3900 non-null   object 
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object 
 12  Shipping Type           3900 non-null   object 
 13  Discount Applied        3900 non-null   object 
 14  Promo Code Used         3900 non-null   object 
 15  Previous Purchases      3900 non-null   int64  
 16  Payment Method          3900 non-null   object 
 17  Frequency of Purchases  3900 non-null   object 
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```



```
df.describe()
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

- **Missing Data Handling:** Checked for null values and imputed missing values in the Review Rating column using the median rating of each product category sold in the particular location.

- **Column Standardization:** Renamed columns to snake case for better readability and documentation.

- **Feature Engineering:**
  - Created age_group column by binding customer ages.
  - Created purchase_frequency_days column from purchase data.

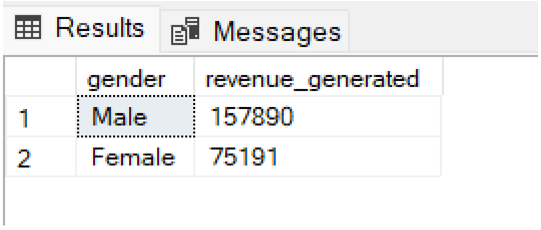
- **Data Consistency Check:** Verified if discount_applied and promo_code_used were redundant, dropped promo_code_used.

- **Database Integration:** Connected python script to SQL-Server and loaded the dataframe into the database for SQL analysis.

4. Data Analysis using SQL

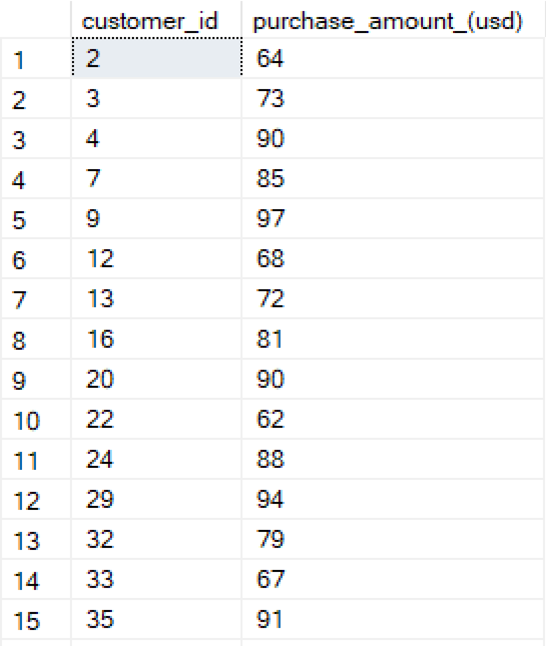
We performed the structure analysis in SQL-Server to answer key business questions.

- Revenue by Gender: Compared total revenue generated by male vs female customers.

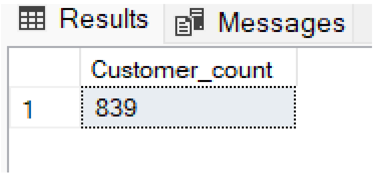


	gender	revenue_generated
1	Male	157890
2	Female	75191

- High Spending Discount Users: Identified customer who used discounts but still spent above average purchase amount.

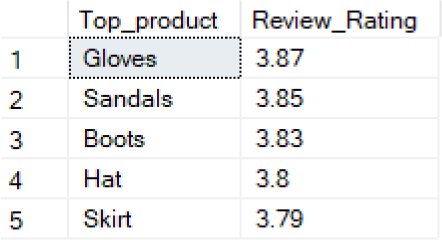


	customer_id	purchase_amount_(usd)
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62
11	24	88
12	29	94
13	32	79
14	33	67
15	35	91



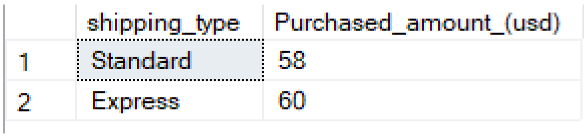
	Customer_count
1	839

- Top 5 Products by Rating: Found products with the highest average review ratings.



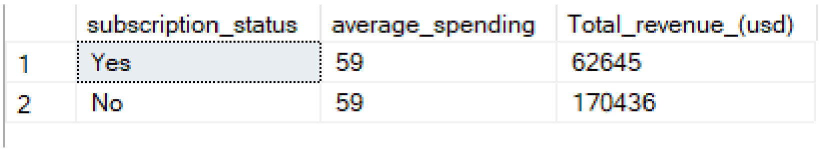
	Top_product	Review_Rating
1	Gloves	3.87
2	Sandals	3.85
3	Boots	3.83
4	Hat	3.8
5	Skirt	3.79

- Shipping Type Comparison: Compared average purchase purchase amounts between Standard and Express Shipping.



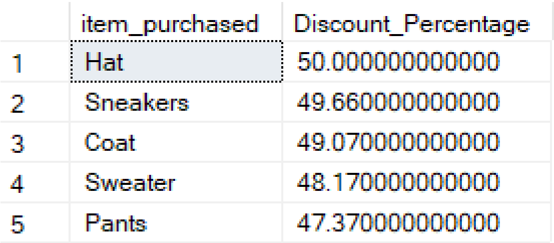
	shipping_type	Purchased_amount_(usd)
1	Standard	58
2	Express	60

- Subscriber vs Non-Subscribers: Compared average spend and total revenue across subscription status.



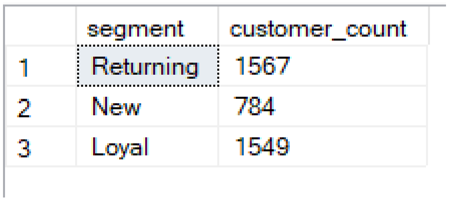
	subscription_status	average_spending	Total_revenue_(usd)
1	Yes	59	62645
2	No	59	170436

- Discount Dependent Products: Identified 5 products with the highest percentage of discount purchases.



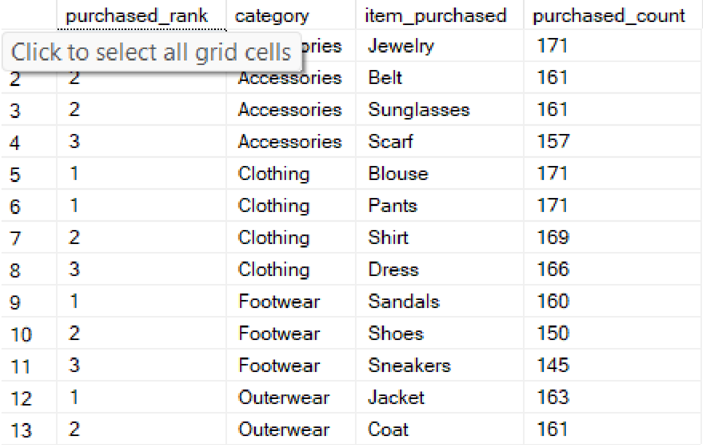
	item_purchased	Discount_Percentage
1	Hat	50.000000000000
2	Sneakers	49.660000000000
3	Coat	49.070000000000
4	Sweater	48.170000000000
5	Pants	47.370000000000

- Customer Segmentation: Classified customers into New, Returning and Loyal segments based on purchase history.



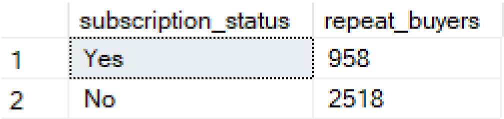
	segment	customer_count
1	Returning	1567
2	New	784
3	Loyal	1549

- Top 3 Products per Category: Listed the most products with in each category.



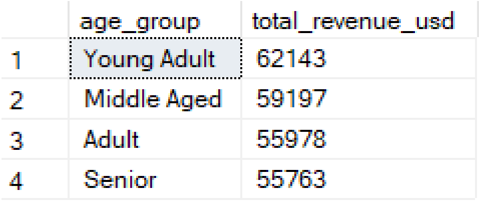
	purchased_rank	category	item_purchased	purchased_count
[illegible]	[illegible]	ries	Jewelry	171
2	2	Accessories	Belt	161
3	2	Accessories	Sunglasses	161
4	3	Accessories	Scarf	157
5	1	Clothing	Blouse	171
6	1	Clothing	Pants	171
7	2	Clothing	Shirt	169
8	3	Clothing	Dress	166
9	1	Footwear	Sandals	160
10	2	Footwear	Shoes	150
11	3	Footwear	Sneakers	145
12	1	Outerwear	Jacket	163
13	2	Outerwear	Coat	161

- Repeat Buyers & Subscriptions: Checked whether the customers with >5 purchases are more likely to subscribe.



	subscription_status	repeat_buyers
1	Yes	958
2	No	2518

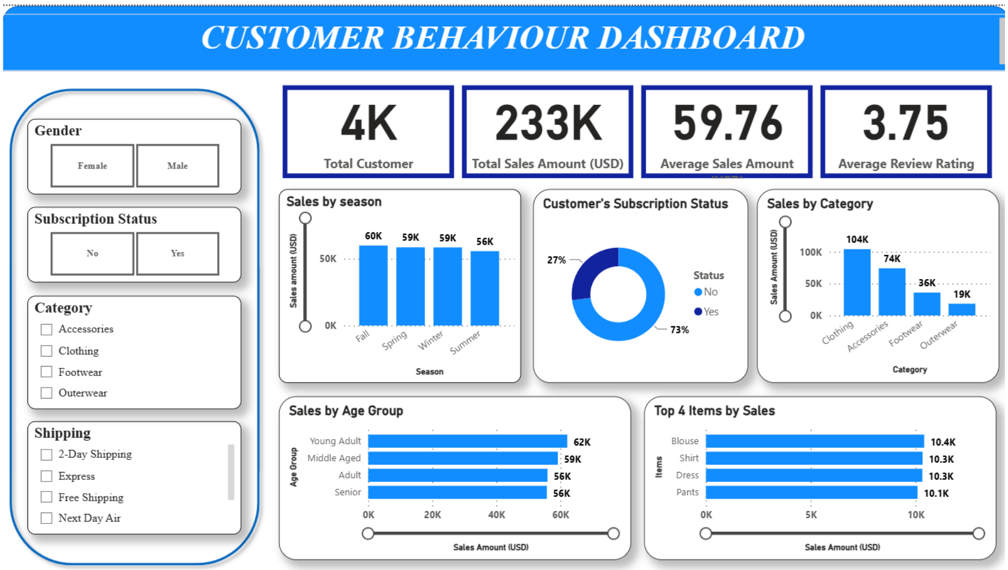
- Revenue by Age Group: Calculate total revenue contribution of each group.



	age_group	total_revenue_usd
1	Young Adult	62143
2	Middle Aged	59197
3	Adult	55978
4	Senior	55763

5. Dashboard in Power BI

Finally built an interactive Power BI dashboard to present insights visually.



6. Business Recommendations

- **Boost Subscription:** Promote exclusive benefits for subscribers.

- **Customer Loyalty Programs:** Reward repeat buyers to move them into the Loyal Segment.

- **Review Discount Policy:** Balance sales boosts with margin control.

- **Product Positioning:** Highlight top rated and best-selling products in segments.

- **Targeted Marketing:** Focus efforts high revenue age groups and express shipping users.